More than interobserver agreement is required for categorization systems comparisons.

Authors:

1. Gloria Palazuelos, MD
   Email: palazuelosgloria@icloud.com
   Telephone: 57+ 6030303 ext 5396
   Fundación Santa Fe de Bogotá
   Institutional address: 116 street # 9-02, Bogotá, Colombia. ZIP: 110111
   First Author

2. Sergio Alfonso Valencia, MD
   Email: sevava92@gmail.com
   Telephone: 57+ 6030303
   Fundación Santa Fe de Bogotá
   Institutional address: 116 street # 9-02 Bogotá, Colombia.
   Corresponding Author

1. Javier Andres Romero, MD
   Email: javier.romero@fsfb.org.co
   Telephone: 57+ 6030303
   Fundación Santa Fe de Bogotá
   Institutional address: 116 street # 9-02 Bogotá, Colombia.

Running title: Interobserver agreement is not enough
Letter to editor

More than interobserver agreement is required for categorization systems comparisons.

Key words: Breast cancer screening, ABCS-K, BI-RADS, Agreement, Kappa

We read with interest the article by Choi, et al., titled “Interobserver agreement in breast ultrasound categorization in the Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness (MUST-BE) trial: results of a preliminary study” (1) in the last issue of Ultrasonography. The article evaluates the interobserver agreement of the Modified Categorization System established by the Alliance for Breast Cancer Screening in Korea (ABCS-K) and compares the results with the BI-RADS categorization. Because present data consists of preliminary results, it is crucial for us to clarify some points.

The authors use the kappa statistic to evaluate interobserver concordance, but they never present a frequency table by categories in each categorization system. The kappa statistic has limitations according to the prevalence. It is known as the kappa paradox, and if there are doubts about its presence, some other statistics can be used to determine concordance (2).

It is interesting to see the good interobserver concordance of the re-modified ABCS-K categorization, but the interobserver concordance of the BI-RADS categorization differs from previous reports k-value of 0.495 vs. 0.51 – 0.53 (3, 4), especially in BI-RADS category 5, k-value of 0.45 vs. 0.71(1, 4). The authors should determine why these differences in the BI-RADS concordance are occurring and should take into account that it could be due to the expertise of the radiologist. It would be interesting as well, to see a table that compares the k-value of BI-RADS categorization by years of experience of the radiologist.

Another subject to review is the methodology used in ABCS-K categorization, because it categorized according to major and minor findings, in contrast to BI-RADS that used the positive predictive value (PPV) of each finding, this difference can be troubleshooting especially in the subcategories 4 (a,b and c). Some minor
findings in ABCS-K were previously proven to have a high PPV as the presence of calcification in the mass (PPV 84.6 – 100%), echogenic halo (PPV of 66.7%), angular margin (PPV 60%) (5). For this reason, it is essential to compare the diagnostic performance of the ABCS – K vs BI-RADS categorization. Although concordance is important to select a categorization system, the diagnostic performance is a central topic of the categorization system to use. As an example, BI-RADS categorization has a good performance with an area under the receiver operating characteristic curve of 0.708 in the fourth edition and 0.690 in the fifth edition (5).

References.


Author Reply

We thank you for your interest and comments on our article titled, "Interobserver agreement in breast ultrasound categorization in the Mammography and Ultrasonography Study for Breast Cancer Screening Effectiveness (MUST-BE) trial: results of a preliminary study".

At first, there were 63 benign and 62 suspicious lesions on US, and 81 benign and 44 breast cancers on final result in the initially modified categorization. Whereas there were 43 benign and 57 suspicious lesions on US, and 53 54 benign and 47 46 breast cancers on final result in the re-modified categorization.

As you mentioned, the kappa statistics are subject to changes in prevalence (1). We described in the Materials and Methods: The proportion of breast cancers among the test series in this article was and those were not low; proportion in the test series of this article was 35.2% (44/125) in the initially modified categorization and 46.0% (46/100) in the re-modified categorization, respectively. We described in the Materials and Methods and those were not low. Therefore, applying kappa statistics to evaluate for interobserver agreement for screening ultrasound in this article is acceptable. On the other hand, prevalence of breast cancers among the test series for screening mammography in the MUST-BE trial was low (1.2%) (2). So, to avoid the kappa paradox, we applied percent agreement as well as kappa statistics in evaluating interobserver agreement for mammography, which was done as a part of a quality control program in the trial.

Although most radiologists participating in the MUST-BE trial were experienced in breast imaging (mean, 10.1 years) in an academic setting, kappa values in this article were lower than those of other studies (3,4). Our results might be influenced by a larger number of cases and observers than other studies (3,4) because the kappa statistic is dependent on the number of categories and observers, and its value is generally higher if there are fewer categories and observers (1). In spite of the lower interobserver agreement using BI-RADS categorization in this article, we believe that it is acceptable it is acceptable in the real world of clinical practice because the interobserver agreement for dichotomous categories (whether to biopsy or not) was moderate and similar to those of other studies (Table 6 in the manuscript).

Regarding suspicious findings, some minor findings including calcification in the mass and angular margin have been known to have high PPVs. Actually, we segregated the suspicious findings into major and minor findings to distinguish category 4 and 5 lesions in quest of both high reproducibility and convenience based on previous studies (5). However, we did not achieve an acceptable value for interobserver agreement about category 4 subcategorization using the modified categorization system. So, we decided not to apply these criteria for the subcategorization of category 4 in the MUST-BE trial. Instead, we will perform further analysis to classify the major and minor findings for the subcategorization of category 4 after completion of a research database.
including information about breast cancer diagnosis.

Reference